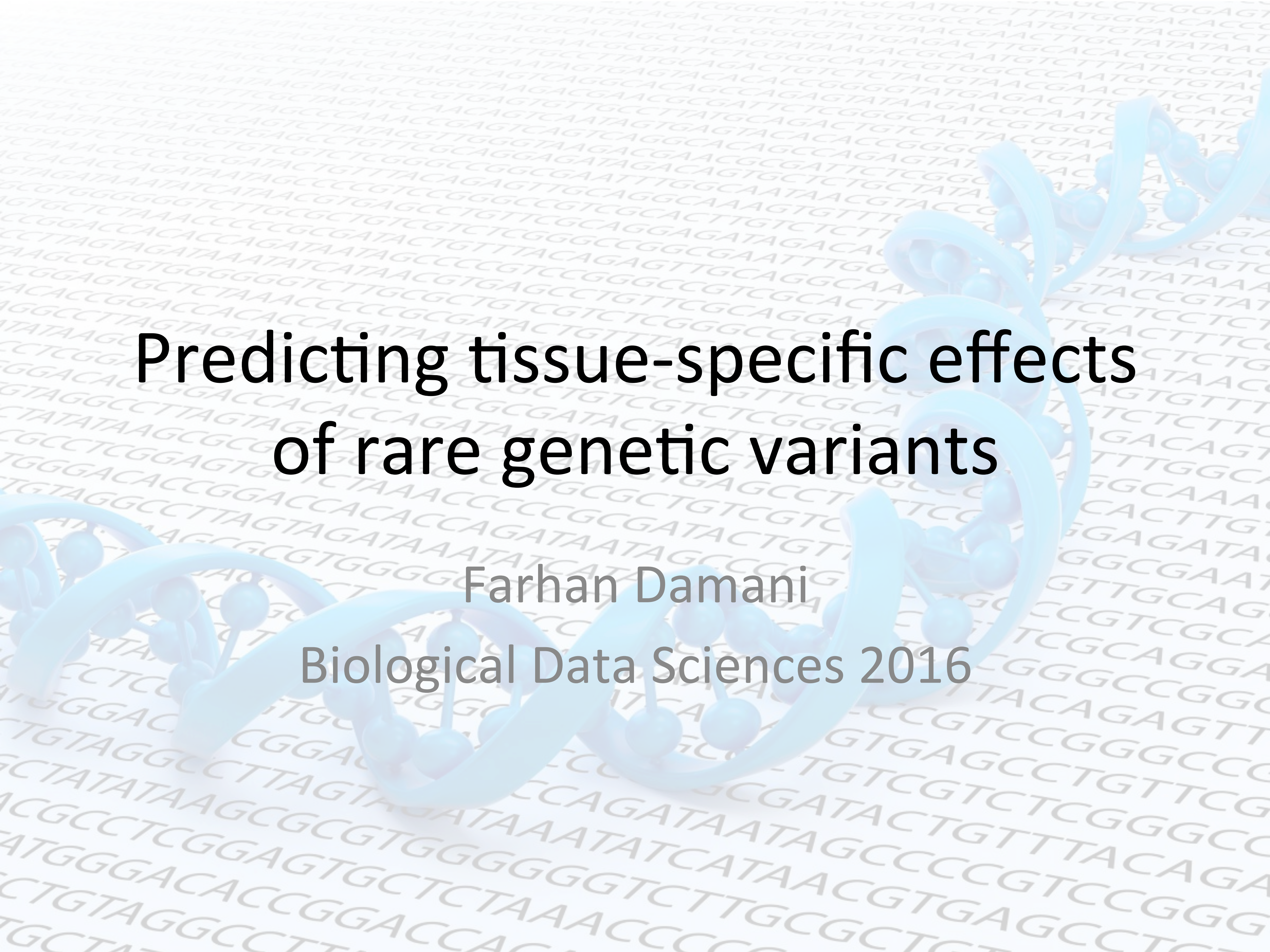


Predicting tissue-specific effects of rare genetic variants

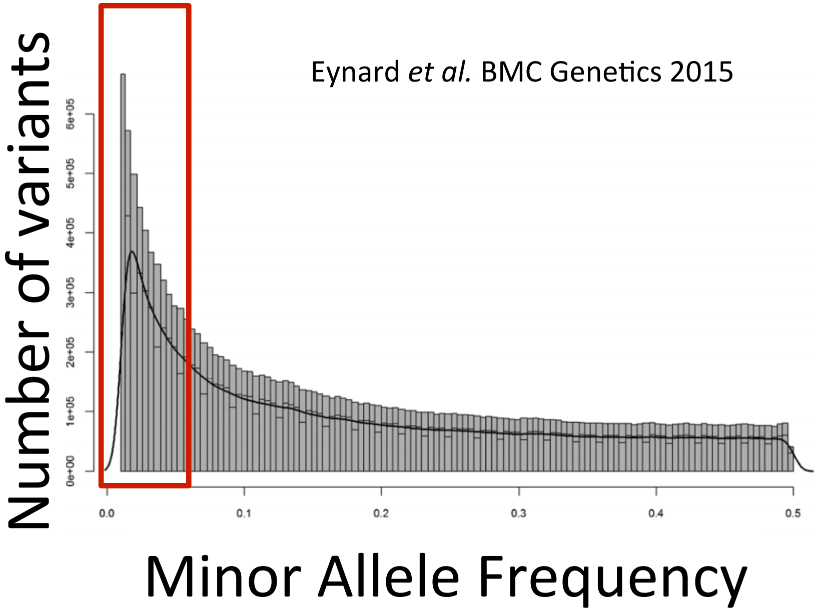
Farhan Damani

Biological Data Sciences 2016

Goal: develop a framework to predict tissue-specific regulatory effects of rare variants

Rare variants are abundant and potentially high-impact

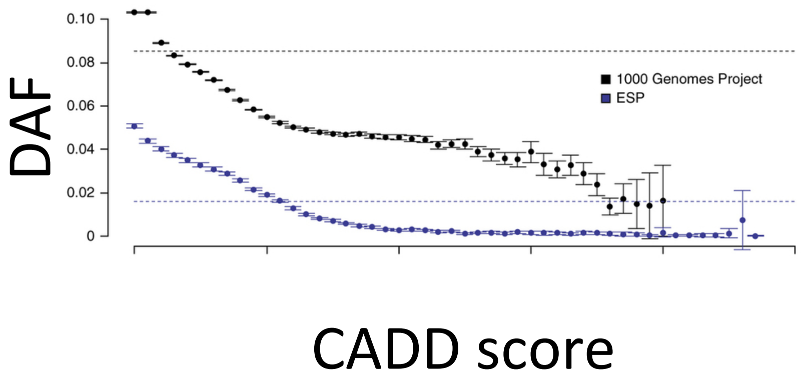
Rare variants defined with minor allele frequency < 1%



Eynard *et al.* BMC Genetics 2015

Enriched for deleterious functional classes

Kircher *et al.* Nature Genetics 2014



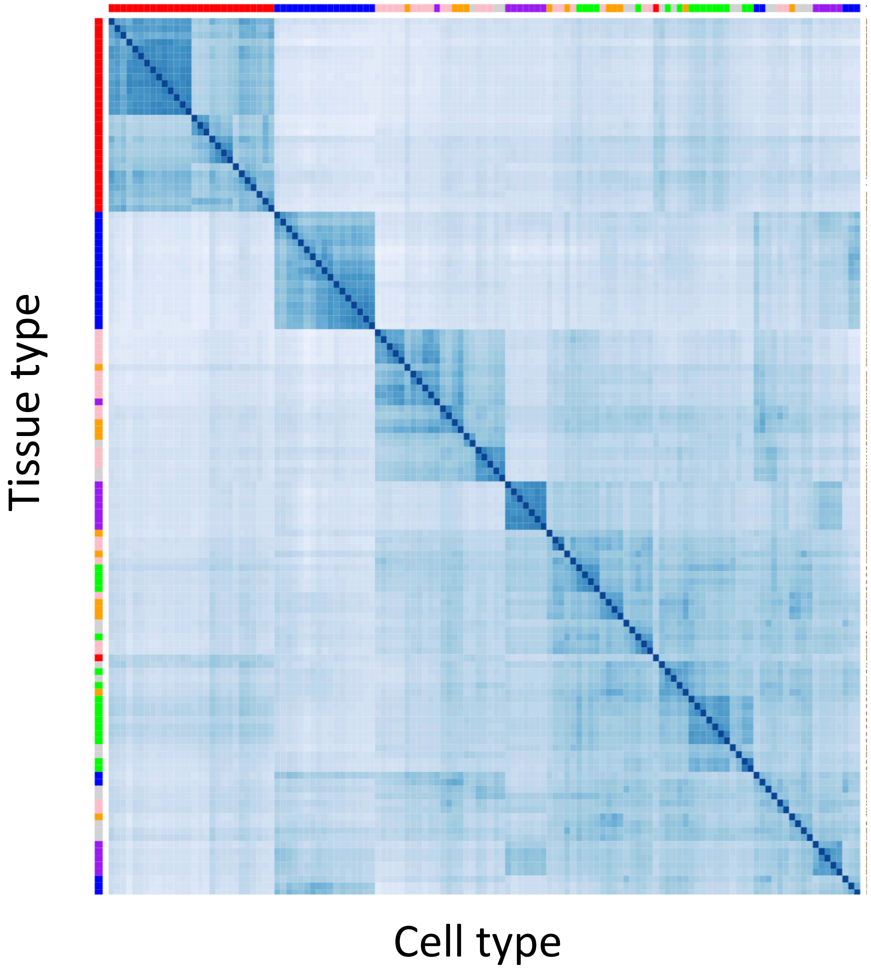
Slide – Alexis Battle

Tissue-specific functionality

- Understanding tissue-specific consequences of noncoding genetic variation is critical to understanding complex traits

Overlap of functional common variants

Backenroth et al. Biorxiv 2016

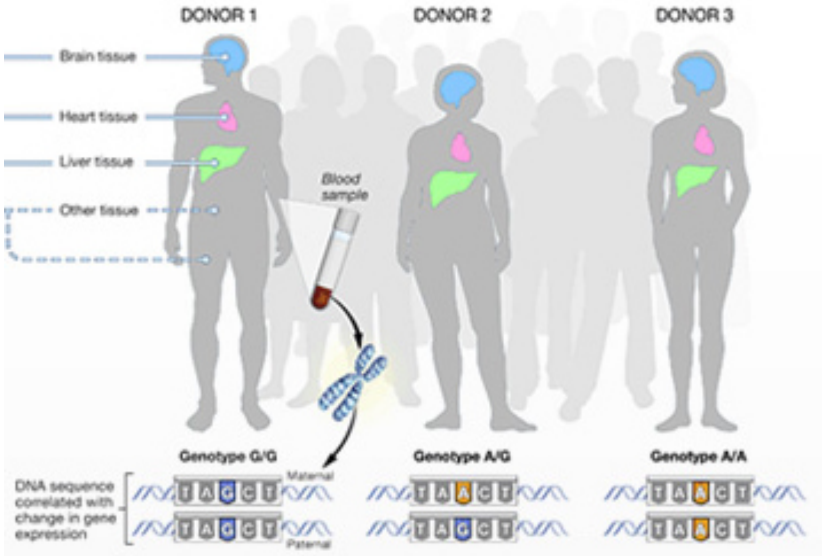


Challenges

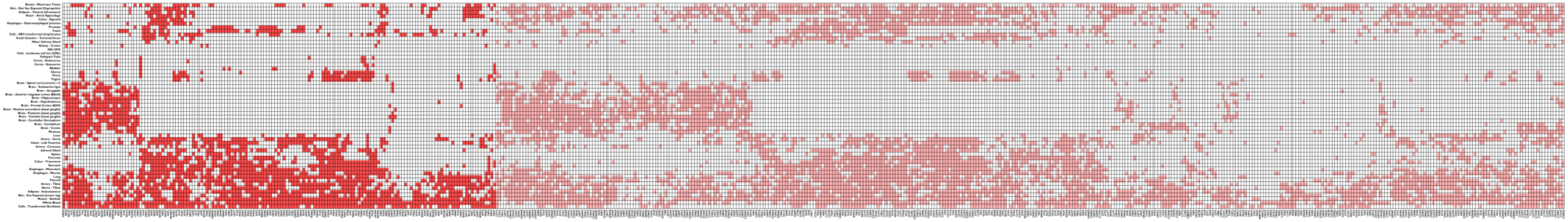
- Even fewer reliable labels in tissue-specific setting
- Each individual tissue has low sample size (RNA-seq)
- Limited samples for each rare SNV

GTEx Project Data

- WGS from 148 donors
  - 114 European Ancestry used here
- 8555 RNA-seq samples from
  - <u>44 tissues</u> from 522 donors



44 tissues

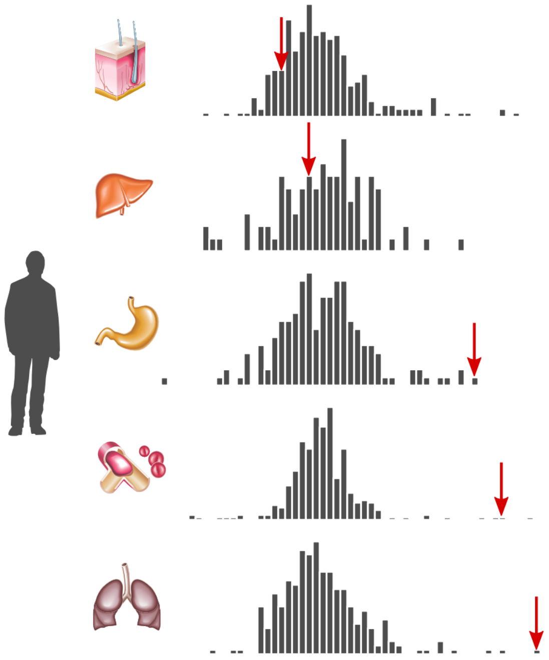


148 individuals (WGS)

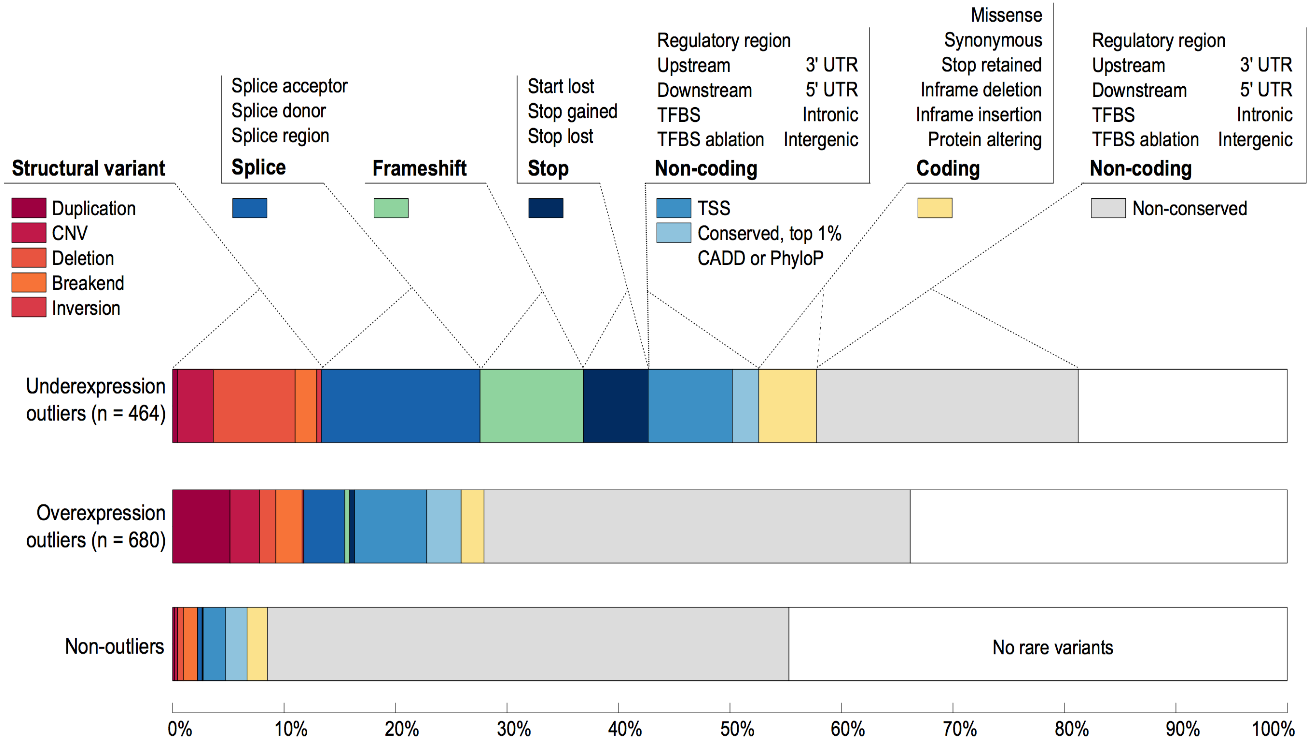
522 individuals (RNA-seq samples)

Expression outliers

What are expression outliers?



Enrichment of functional variants among outliers



Li et al. The impact of rare variation. Biorxiv http://biorxiv.org/content/early/2016/09/09/074443

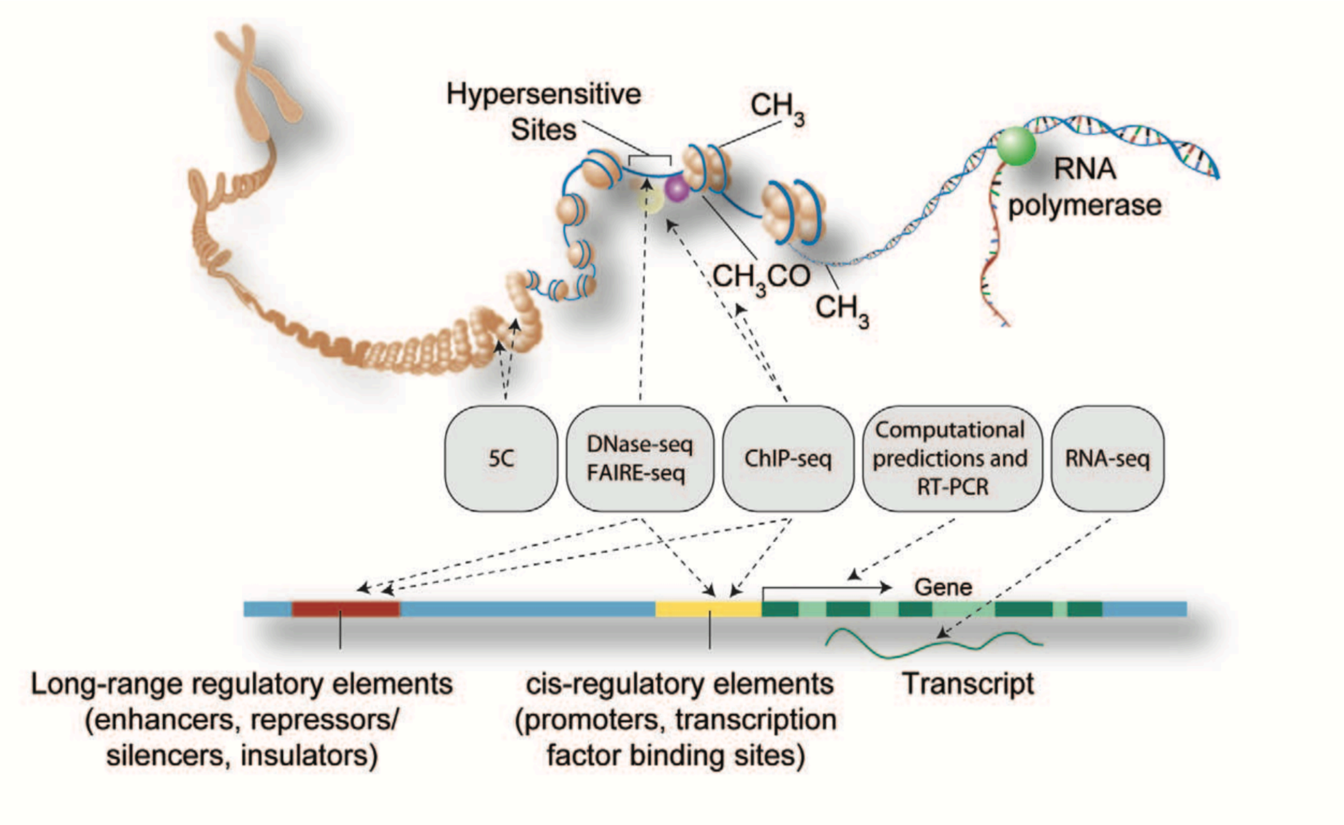
Genomic features

(1) regulatory elements
(2) variant predictor summary statistics

- **Variant effect predictor**
- **CADD**
- **DANN**
- **…**

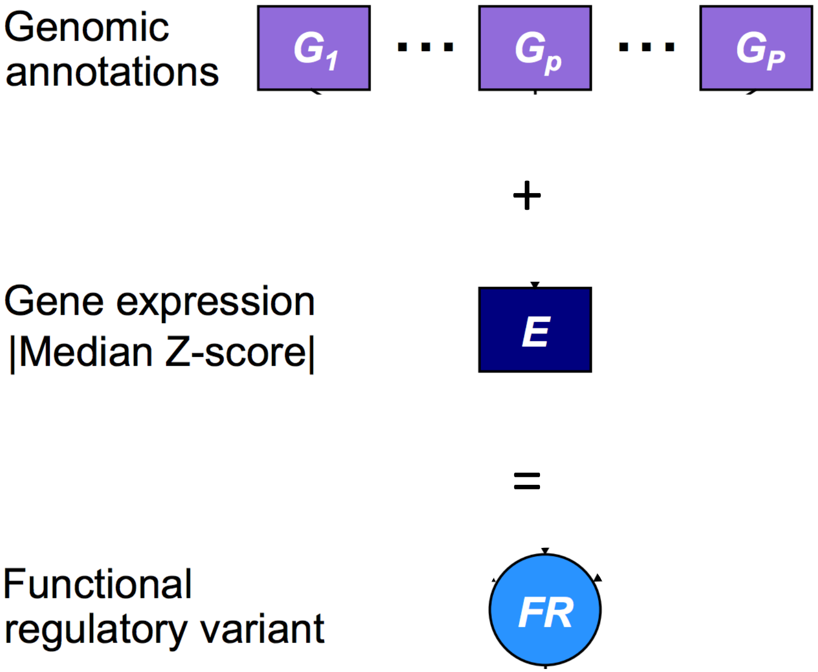
Genomic features

ENCODE Project Consortium. Plos Biology 2011.


Hypersensitive Sites
CH_3
CH_3CO
CH_3
RNA polymerase
5C
DNase-seq FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

- <u>Tissue-specific</u> promoters/ enhancers
- Conservation scores
- Transcription factor binding sites
- CpG sites
- ChromHMM

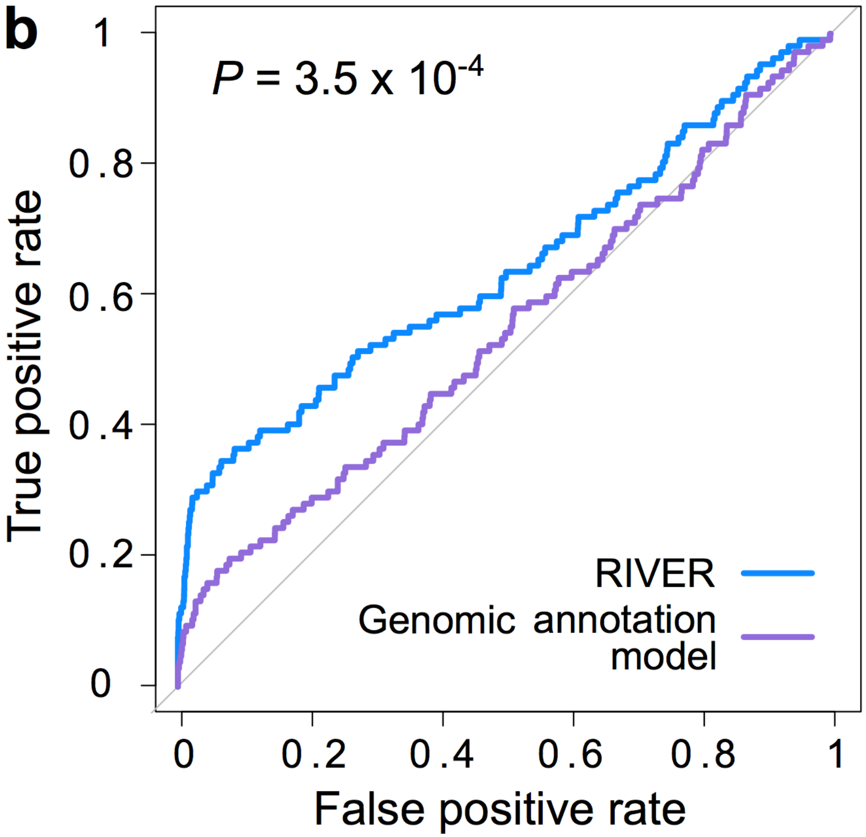
Related work on tissue-shared effects



Genomic annotations

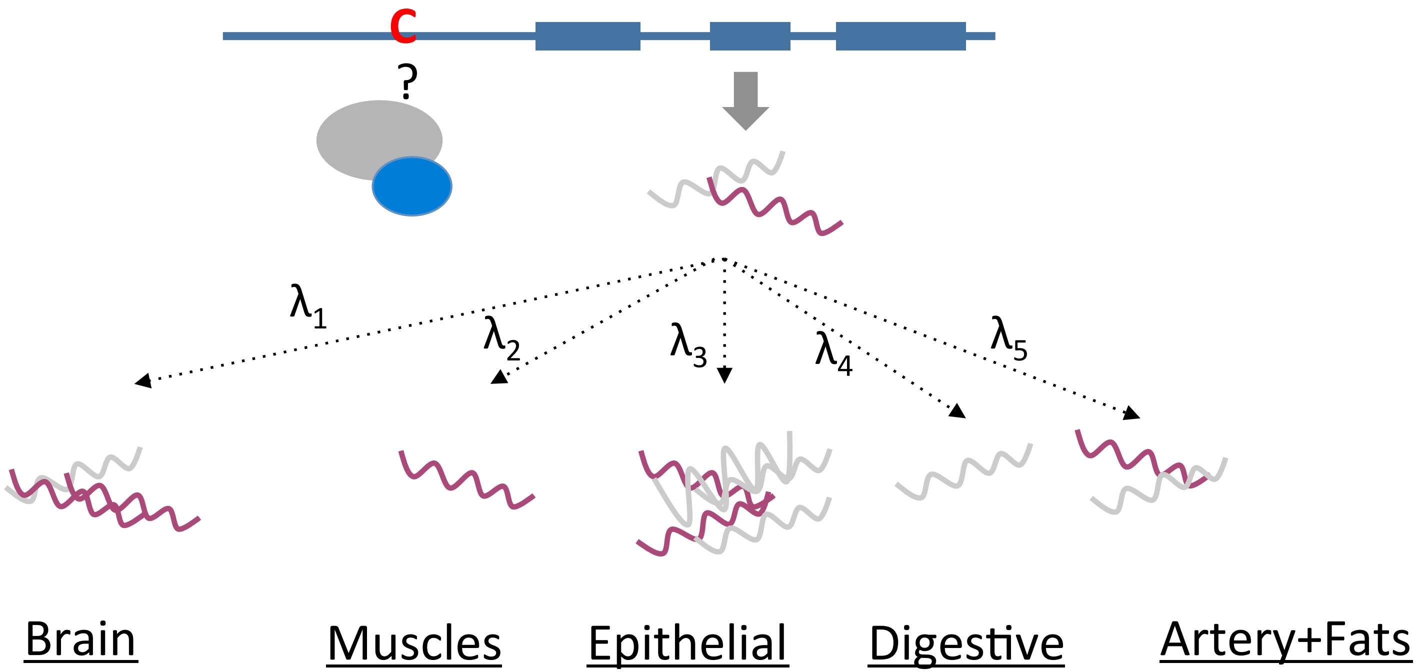
Gene expression |Median Z-score|

Functional regulatory variant

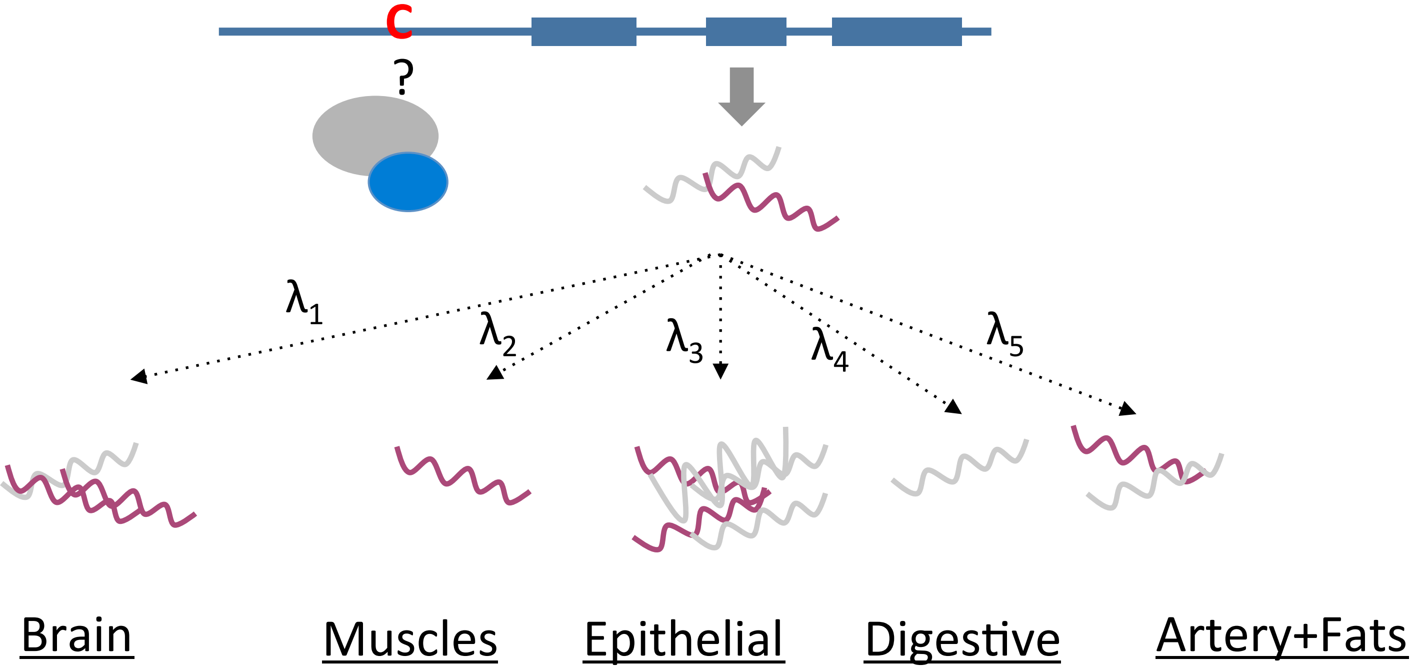


Li et al. The impact of rare variation. Biorxiv http://biorxiv.org/content/early/2016/09/09/074443

Learning tissue-specific effects as individual tasks



Learning tissue-specific effects as individual tasks



Brain Muscles Epithelial Digestive Artery+Fats

Expression outliers are noisier based on smaller sets of tissues

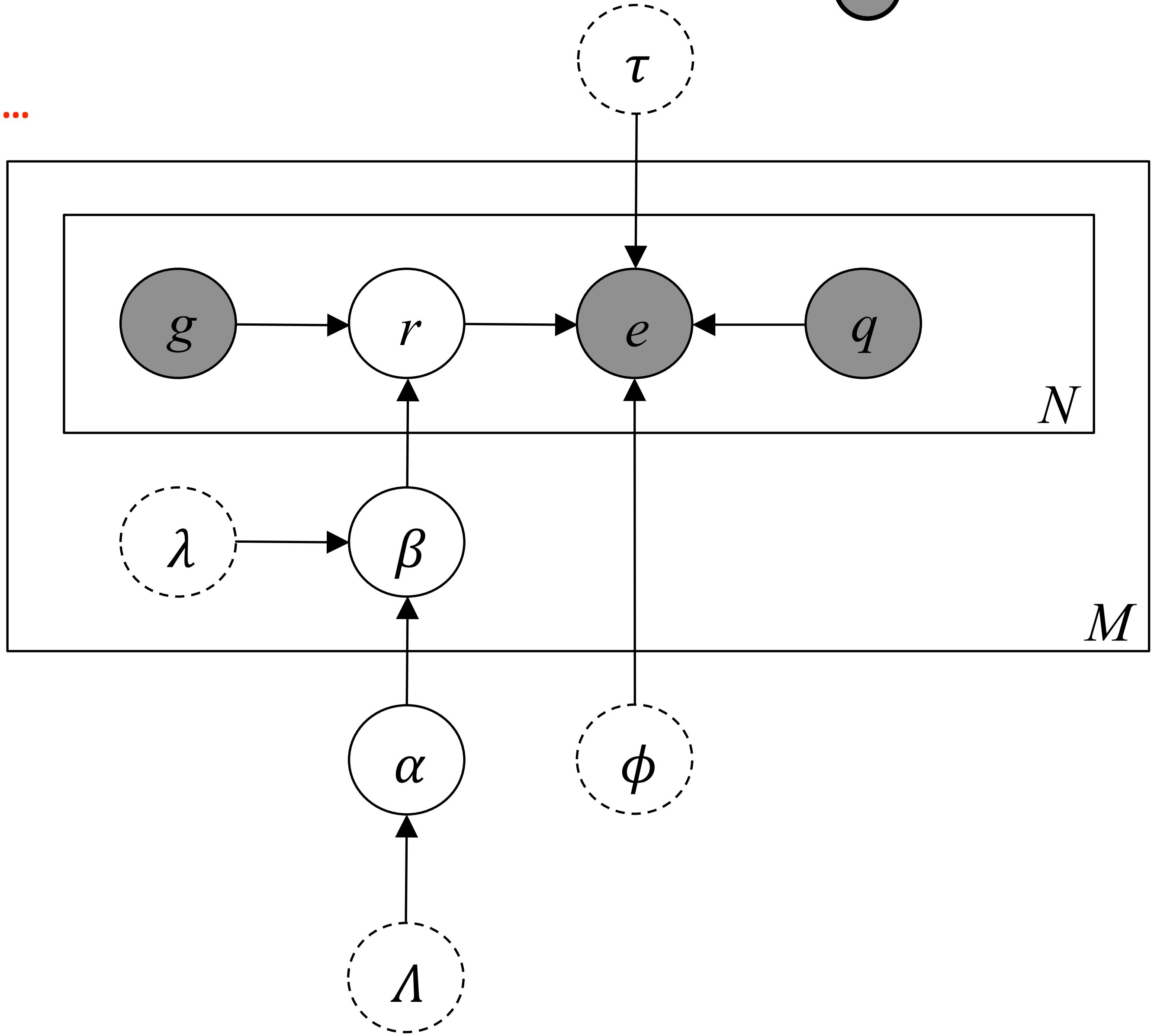
○ unobserved

● observed

Graphical model

Boxes represent replicates…

- **M tissues**
- **N individual by gene samples**

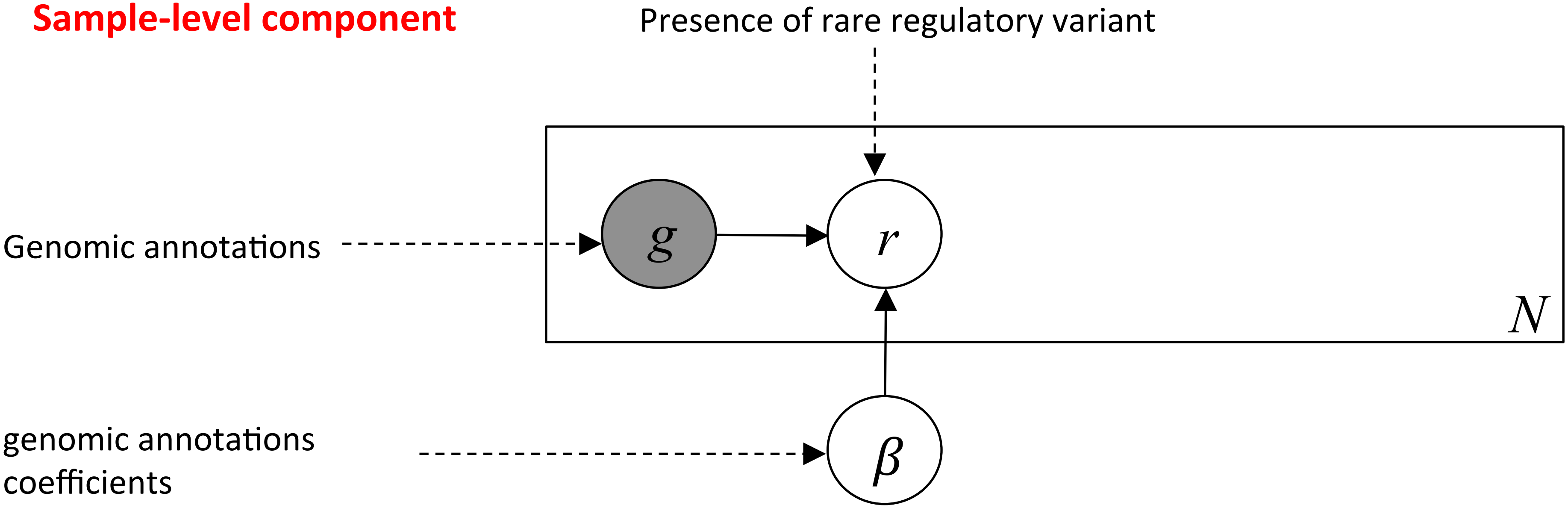


Graphical model

○ unobserved

● observed

Sample-level component



Presence of rare regulatory variant

Genomic annotations

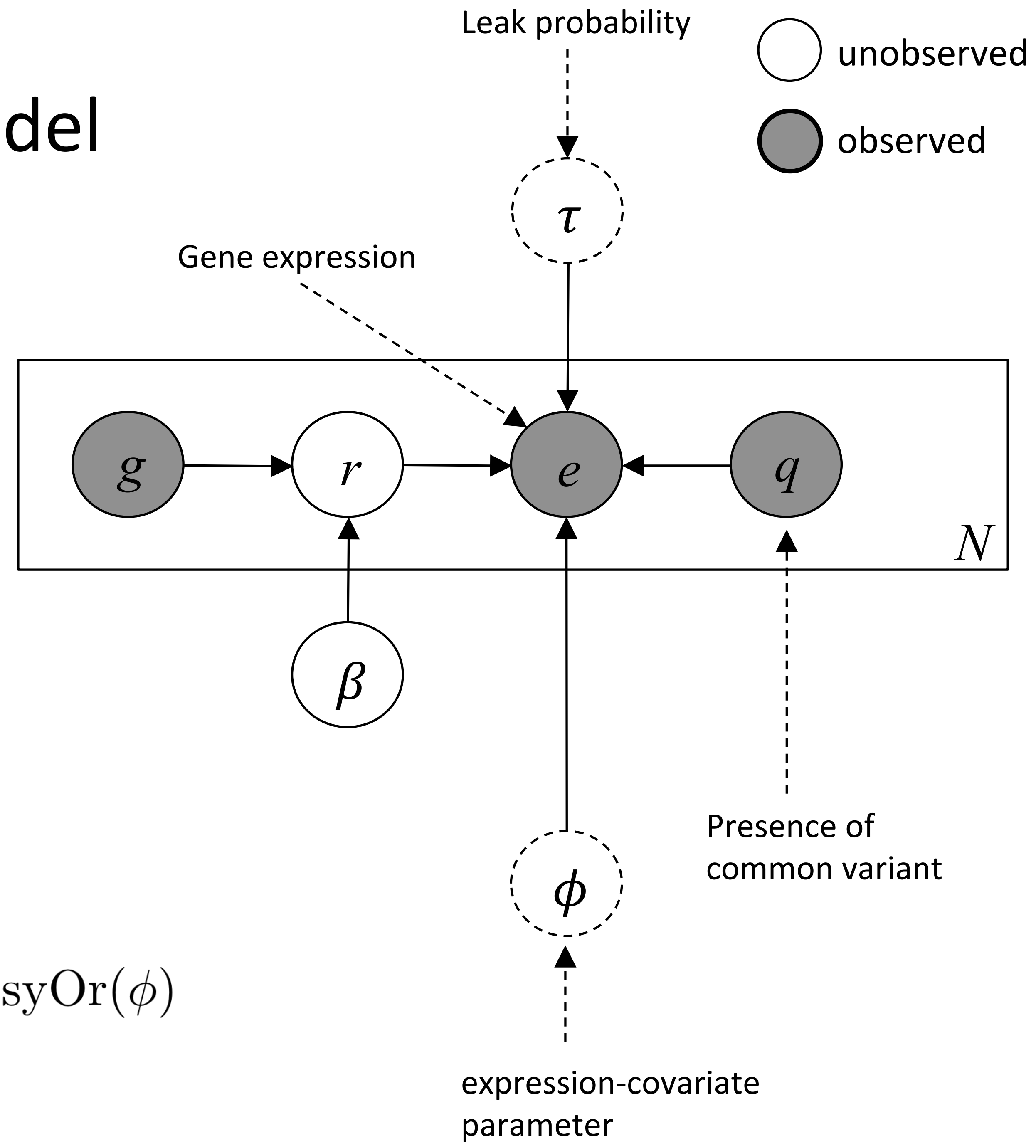
genomic annotations coefficients

$$r_{ci}|g_{ci},\beta_c \sim Bernoulli(logit^{-1}(g_{ci}))$$

$$\psi(g_{ci}) = \frac{1}{1+e^{-\beta_c^T g_{ci}}}$$

Graphical model

Sample-level component



Leak probability

Gene expression

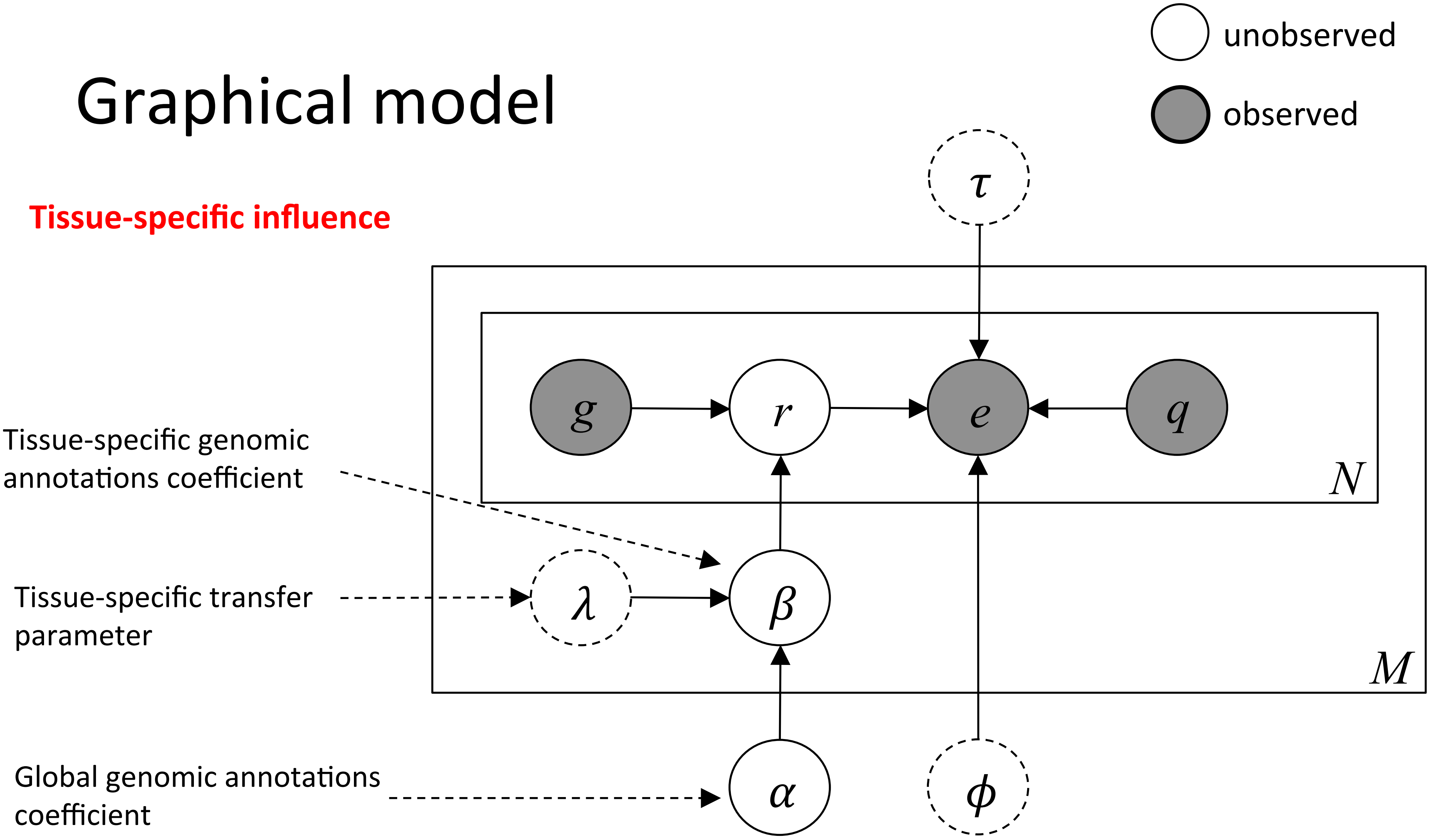
Presence of common variant

expression-covariate parameter

$$e_{ci}|r_{ci}, q_{ci}, \tau, \phi \sim \mathrm{NoisyOr}(\phi)$$

Graphical model

Tissue-specific influence


unobserved
observed
τ
g
r
e
q
N
λ
β
M
α
φ
Tissue-specific genomic annotations coefficient
Tissue-specific transfer parameter
Global genomic annotations coefficient

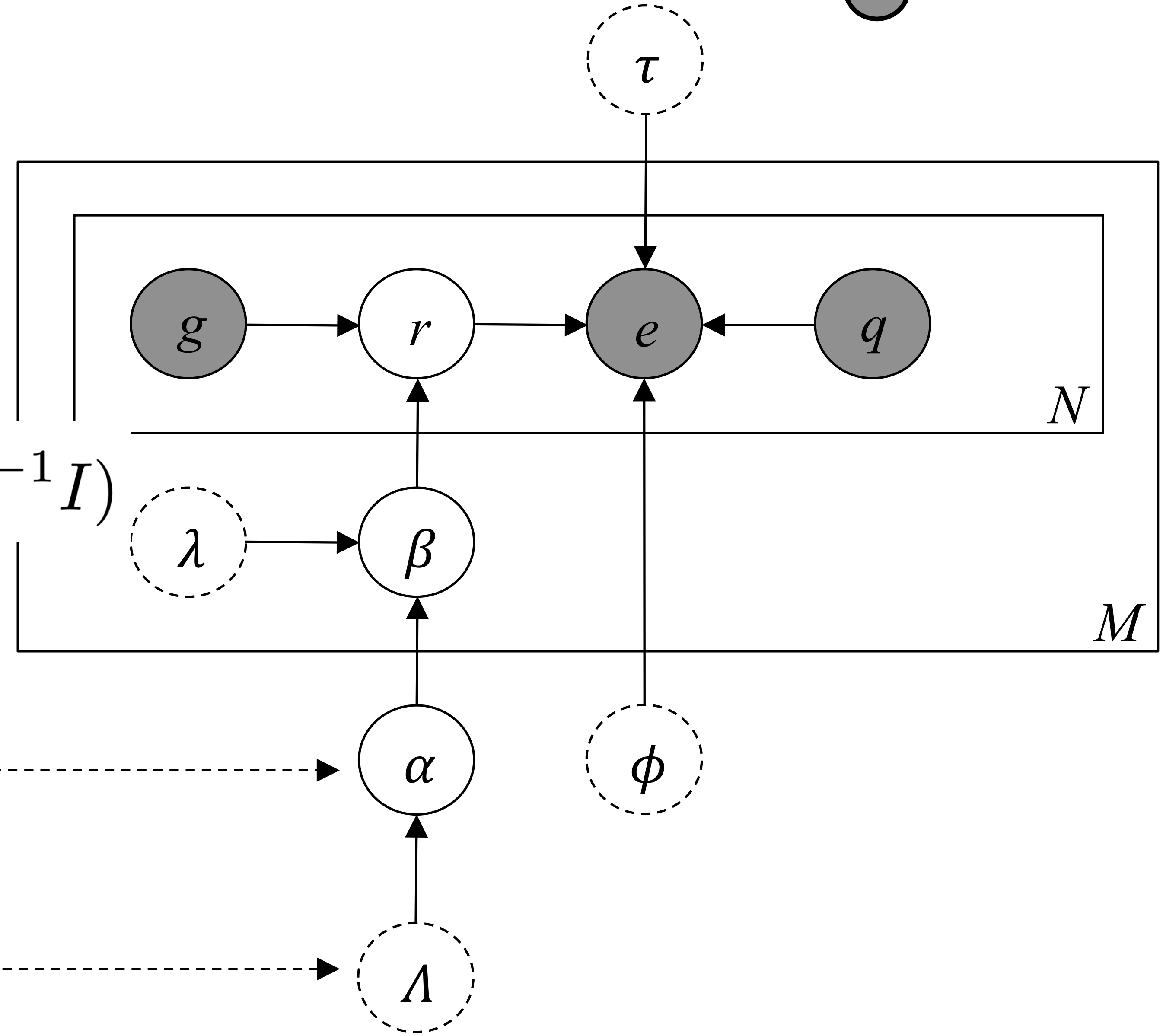
$$\beta_{cj}|\alpha_j, \lambda_c \sim \mathcal{N}(\alpha_j, \lambda_c^{-1})$$

Graphical model

○ unobserved

● observed

Global influence

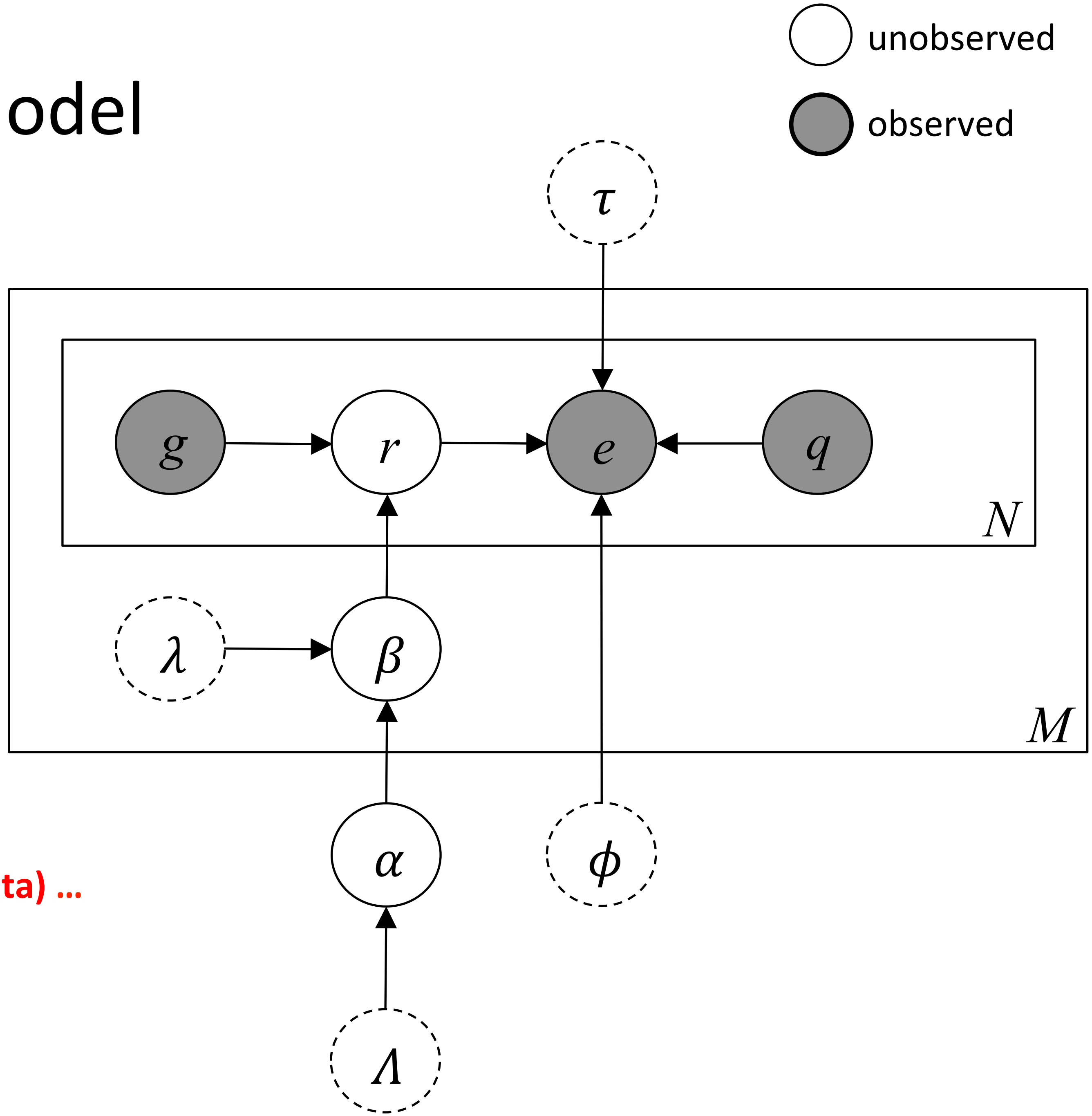


$$\alpha|\Lambda \sim \mathcal{N}(\vec{0}, \Lambda^{-1} I)$$

Global genomic annotations coefficient

Global transfer parameter

Graphical model



We want to infer p(regulatory variant | data) …

Objective function

$$\log p(\boldsymbol{e},\boldsymbol{g},\boldsymbol{r},\boldsymbol{q},\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\tau},\alpha,\Lambda,\phi) = \underbrace{\log p(\alpha|\Lambda)}_{\text{(A) global influence}} + \underbrace{\sum_{c=1}^{M}\left(\sum_{j=1}^{L}\log p(\beta_{cj}|\alpha_j,\lambda_c)\right)}_{\text{(B) tissue-specific influence}}$$

$$+ \underbrace{\sum_{i=1}^{N_c}\log\sum_{r_{ci}} p(e_{ci}|r_{ci},q_{ci},\tau_c,\phi)p(r_{ci}|g_{ci},\beta_c)}_{\text{(C) sample-level component}}$$

Objective function

$$\log p(\boldsymbol{e}, \boldsymbol{g}, \boldsymbol{r}, \boldsymbol{q}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\tau}, \alpha, \Lambda, \phi) = \underbrace{\log p(\alpha|\Lambda)}_{\text{(A) global influence}} + \underbrace{\sum_{c=1}^{M}\left(\sum_{j=1}^{L}\log p(\beta_{cj}|\alpha_j, \lambda_c)\right)}_{\text{(B) tissue-specific influence}}$$

$$+ \underbrace{\sum_{i=1}^{N_c}\log\sum_{r_{ci}} p(e_{ci}|r_{ci}, q_{ci}, \tau_c, \phi)p(r_{ci}|g_{ci}, \beta_c)}_{\text{(C) sample-level component}}$$

$$\Theta = \{\underbrace{\beta_{1_1:T_G}, \phi, \alpha,}_{\text{parameters}} \quad \underbrace{\lambda_{1:T}, \tau, \Lambda}_{\text{hyperparameters}} \quad \}$$

Hyperparameter setting

$$\Theta = \{\underbrace{\beta_{1_1:T_G}, \phi, \alpha,}_{\text{parameters}} \quad \underbrace{\lambda_{1:T}, \tau, \Lambda}_{\text{hyperparameters}} \quad \}$$

- $\{\lambda_{1:T}, \Lambda\}$ (transfer parameters)

  Bootstrap estimation: $\lambda_c{}^{-1} = \sigma_c{}^2 = \dfrac{\sum_{i=1}^{K} \sum_{j=1}^{L} (\beta_{cj}^{(i)} - \alpha_j^{(i)})^2}{(K-1)L}$

- $\{\tau\}$ (leak probability)

  Categorical distribution

Optimizing the objective using EM

$$\Theta = \{\underbrace{\beta_{1_{1:T_G}}, \phi, \alpha,}_{\text{parameters}} \quad \underbrace{\lambda_{1:T}, \tau, \Lambda}_{\text{hyperparameters}} \quad \}$$

- Expectation step
  - Exact inference $\quad q_{ci}(r_{ci}) = p(r_{ci}|\text{data}, \Theta)$
- Maximization Step

$$\text{Coordinate gradient descent} \begin{cases} \alpha_j = \dfrac{\sum_{c=1}^{M} \lambda_c \beta_{cj}}{\Lambda + \sum_{c=1}^{M} \lambda_c} \\ \beta_{cj}^{t+1} = \beta_{cj}^{t} - \nabla f(\beta_{cj}^{t}, \alpha_j^{t}, q_{ci}, g_{ci}) \end{cases}$$

ϕ NoisyOr update

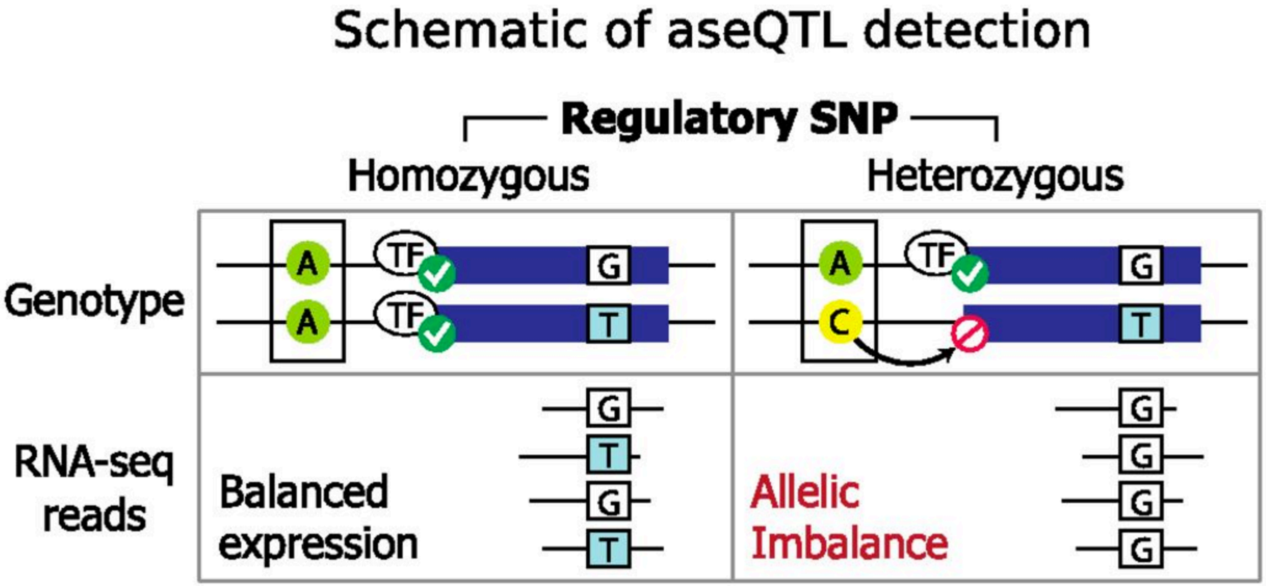
Results

Allelic imbalance presents strong evidence for regulatory variation

Battle et al. Genome Research 2013

Strong evidence of <u>causal cis-regulatory impact</u>

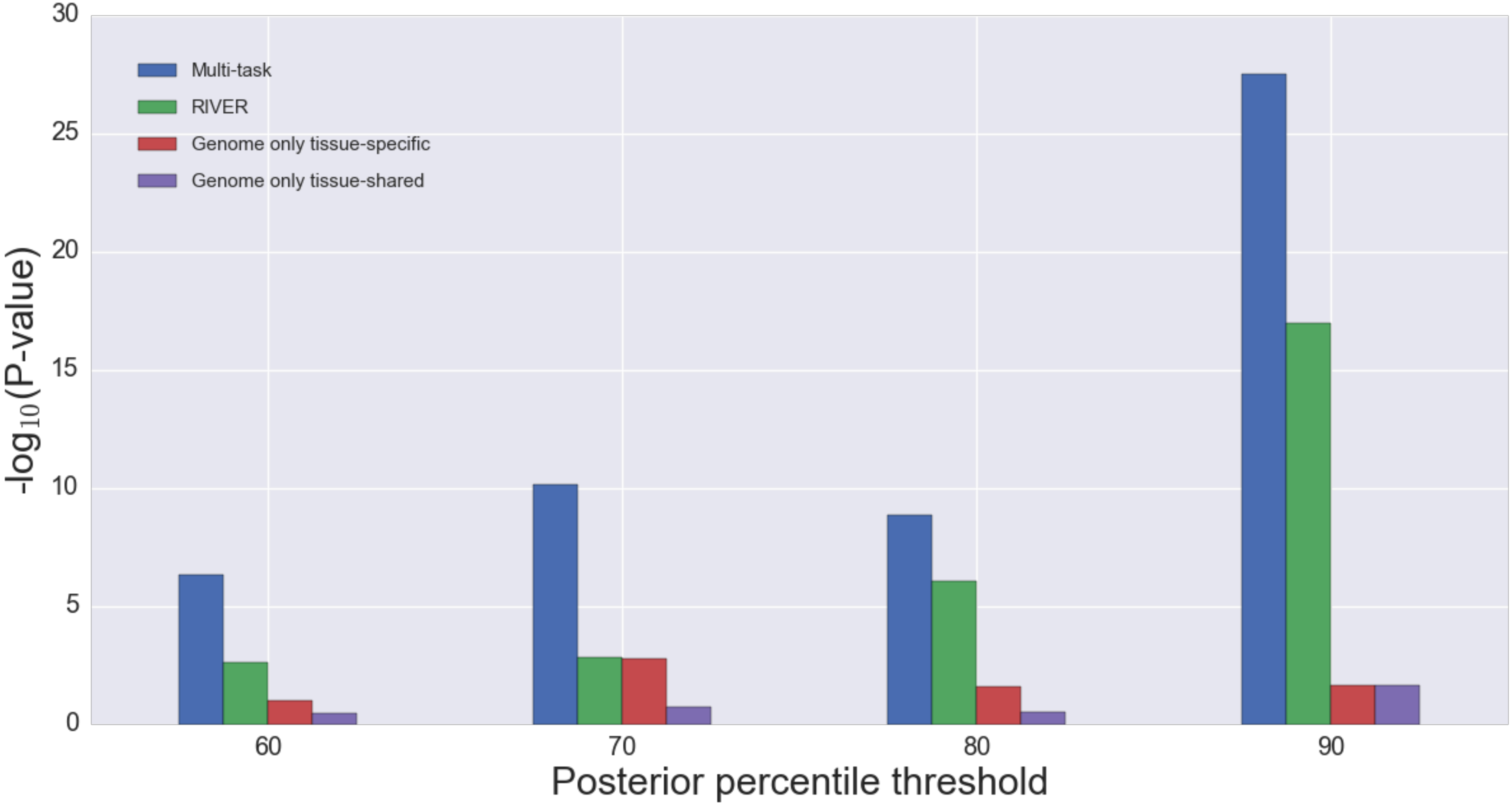
Almost all rare variants in our cohort are heterozygous

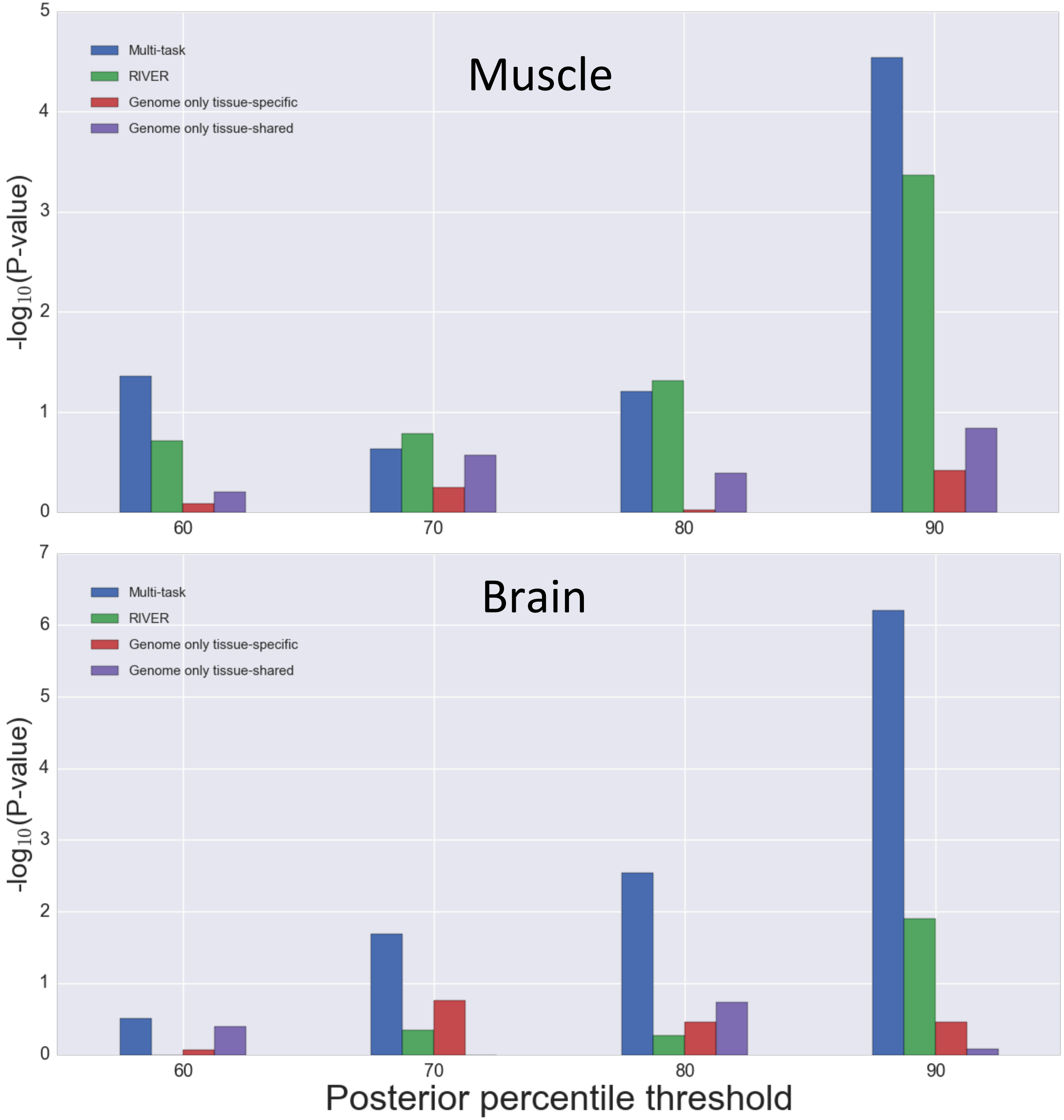


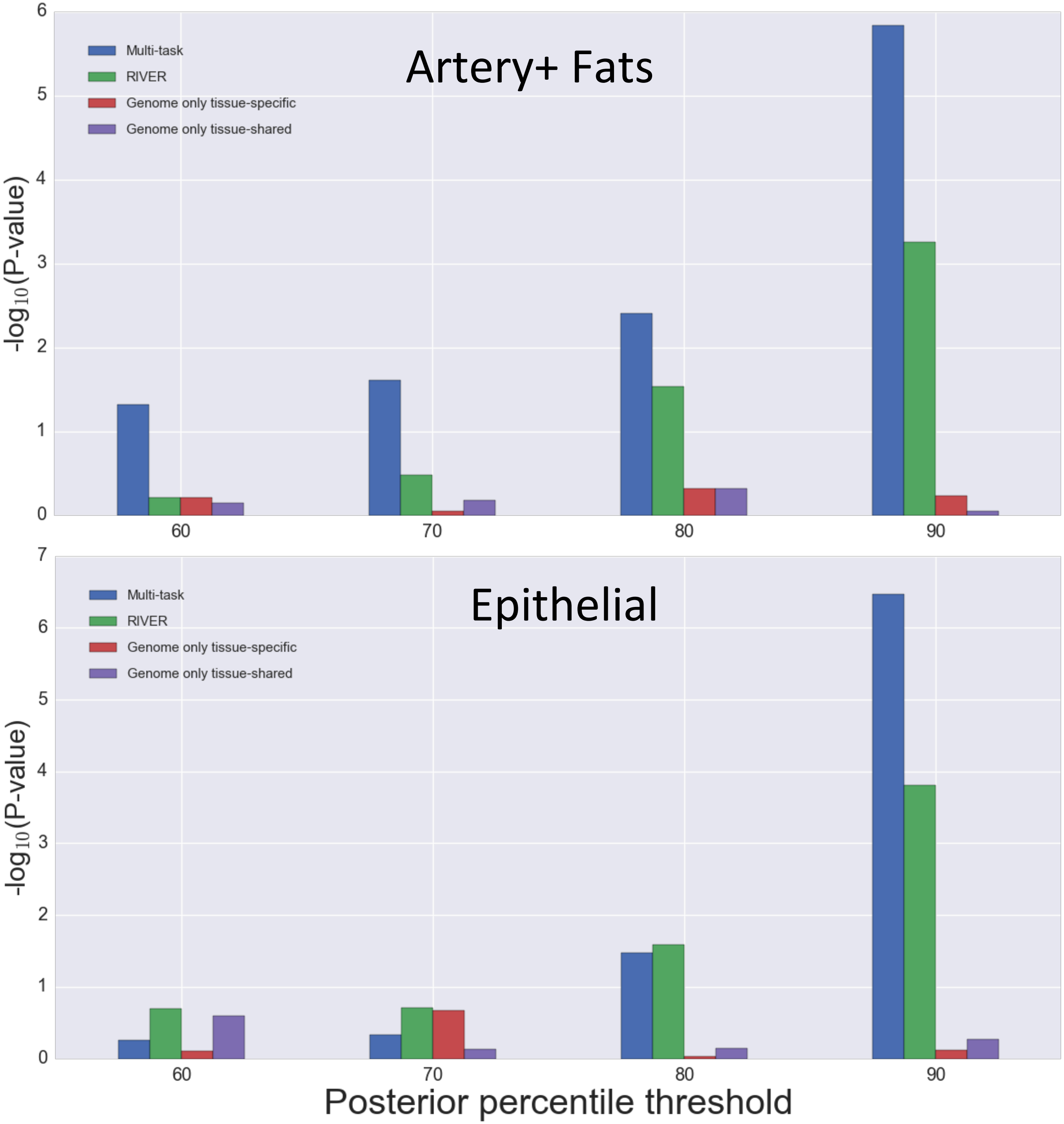
Zhang et al. Nature Methods 2009: “we found that the variation of allelic ratios in gene expression among different cell lines was primarily explained by genetic variations…”

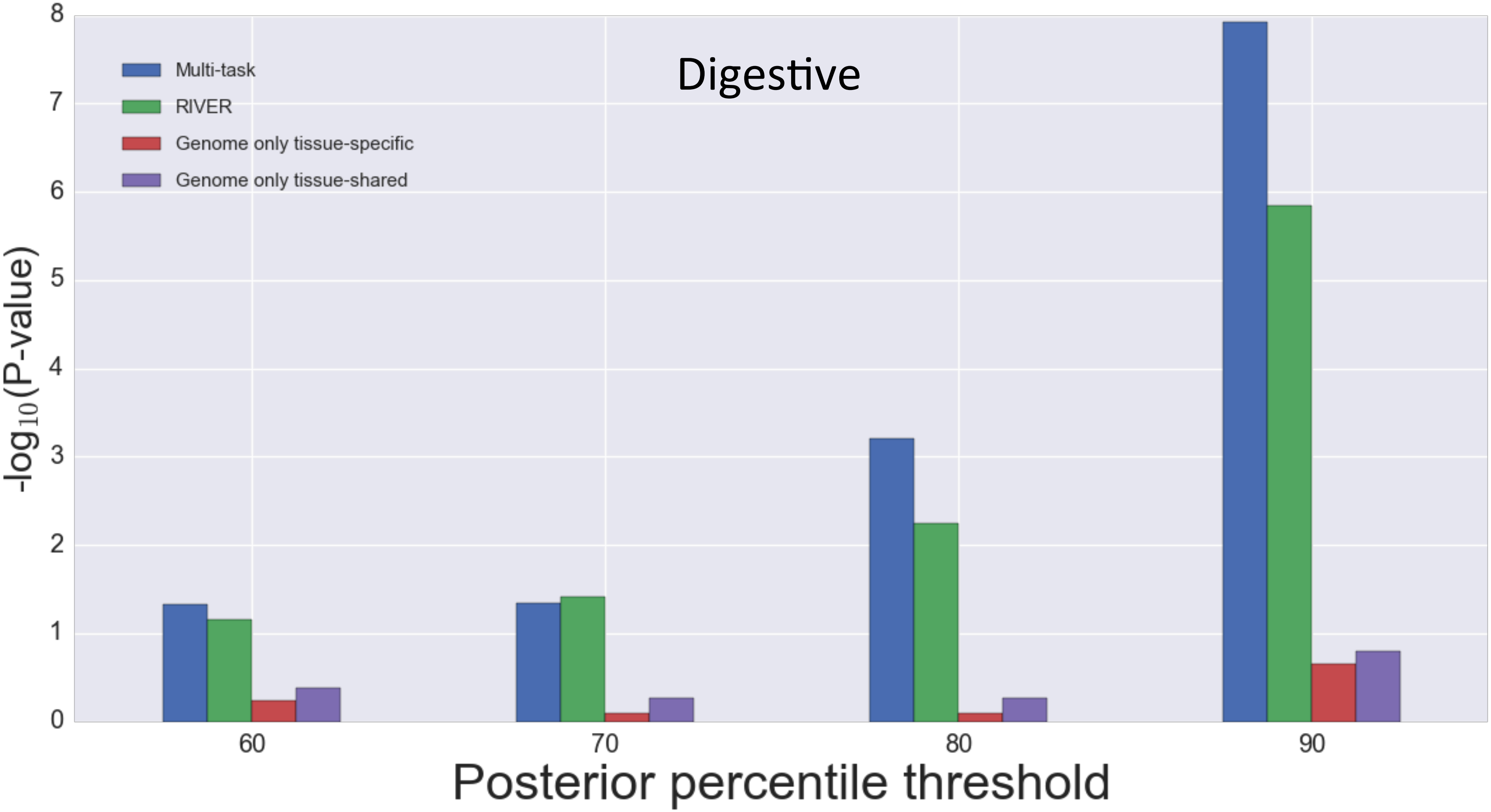
Yan et. al. Science 2002: “We estimated that this approach could confidently identify variations when the differences between expression of the two alleles differed by more than 20%.”

Posteriors are predictive of allelic imbalance

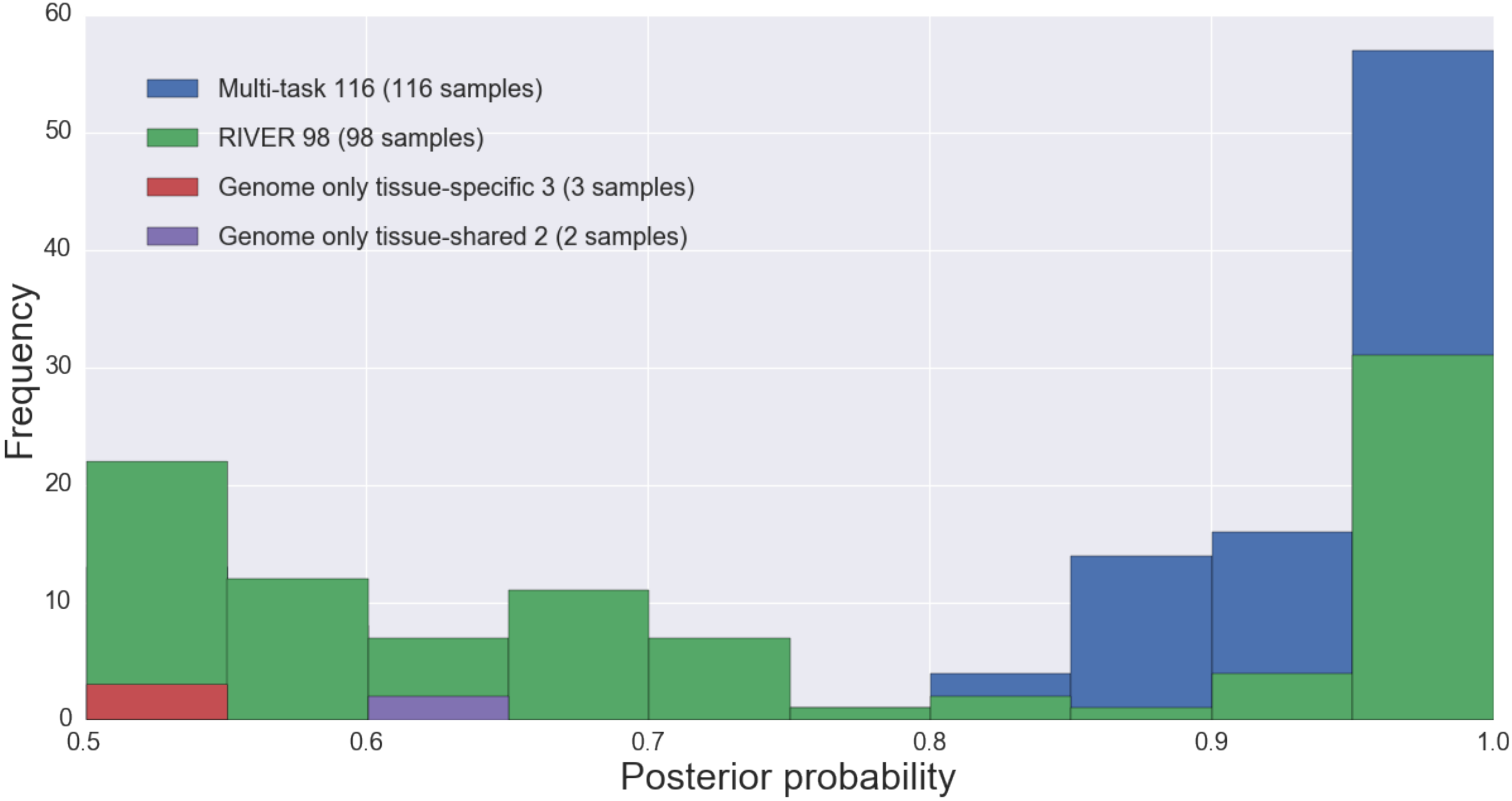



Muscle
Multi-task
RIVER
Genome only tissue-specific
Genome only tissue-shared
-log$_{10}$(P-value)
0
1
2
3
4
5
60
70
80
90
Brain
Multi-task
RIVER
Genome only tissue-specific
Genome only tissue-shared
-log$_{10}$(P-value)
0
1
2
3
4
5
6
7
60
70
80
90
Posterior percentile threshold

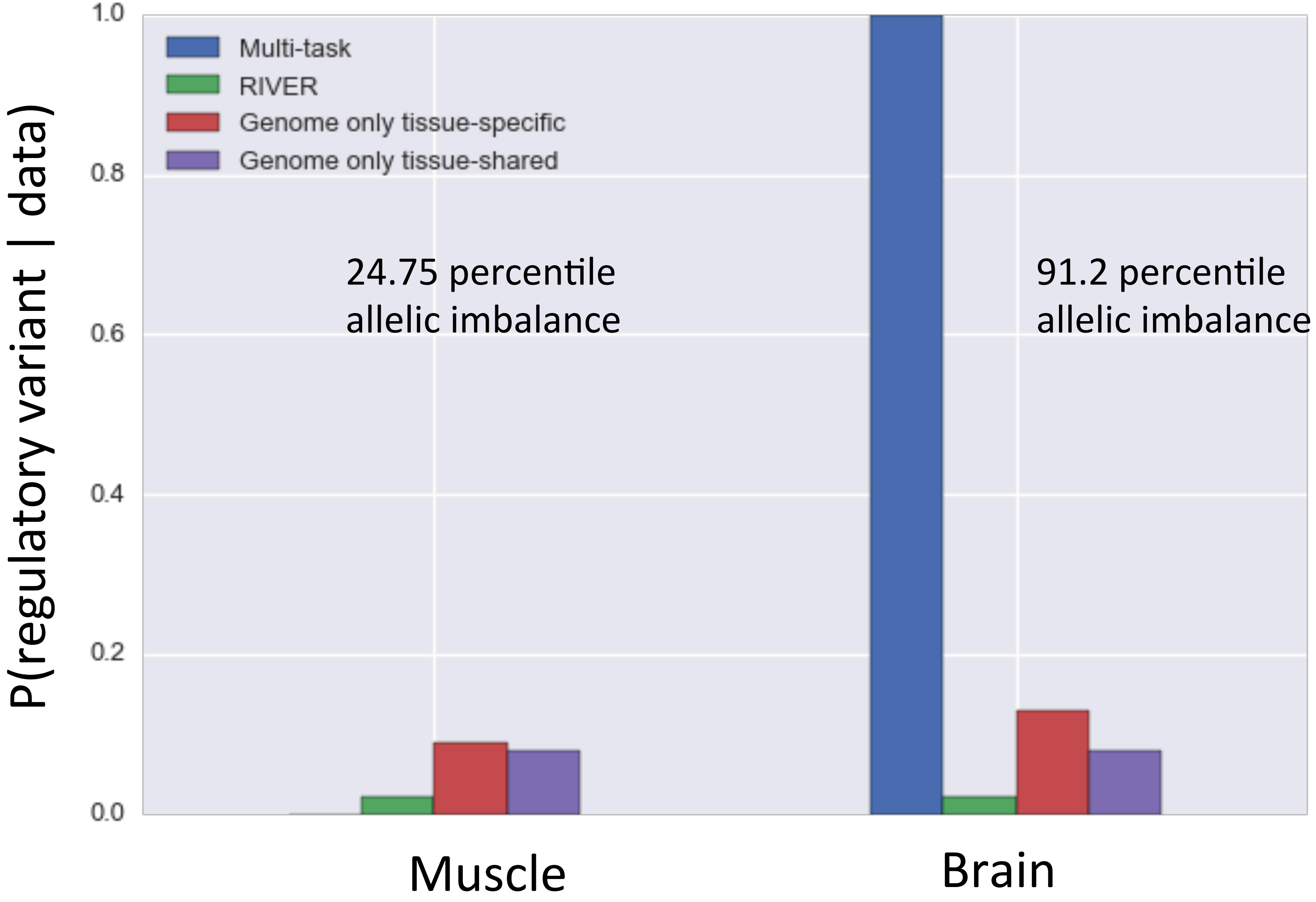

Artery+ Fats
Multi-task
RIVER
Genome only tissue-specific
Genome only tissue-shared
$-\log_{10}$(P-value)
0
1
2
3
4
5
6
60
70
80
90
Epithelial
Multi-task
RIVER
Genome only tissue-specific
Genome only tissue-shared
$-\log_{10}$(P-value)
0
1
2
3
4
5
6
7
60
70
80
90
Posterior percentile threshold


Digestive
Multi-task
RIVER
Genome only tissue-specific
Genome only tissue-shared
-log$_{10}$(P-value)
0
1
2
3
4
5
6
7
8
60
70
80
90
Posterior percentile threshold

Our predictions are also confident



Rare regulatory variant nearby GCAT



24.75 percentile allelic imbalance

91.2 percentile allelic imbalance

Conclusion

We developed a framework for regulatory rare variant prediction

We compared our predictions to measured allelic imbalance

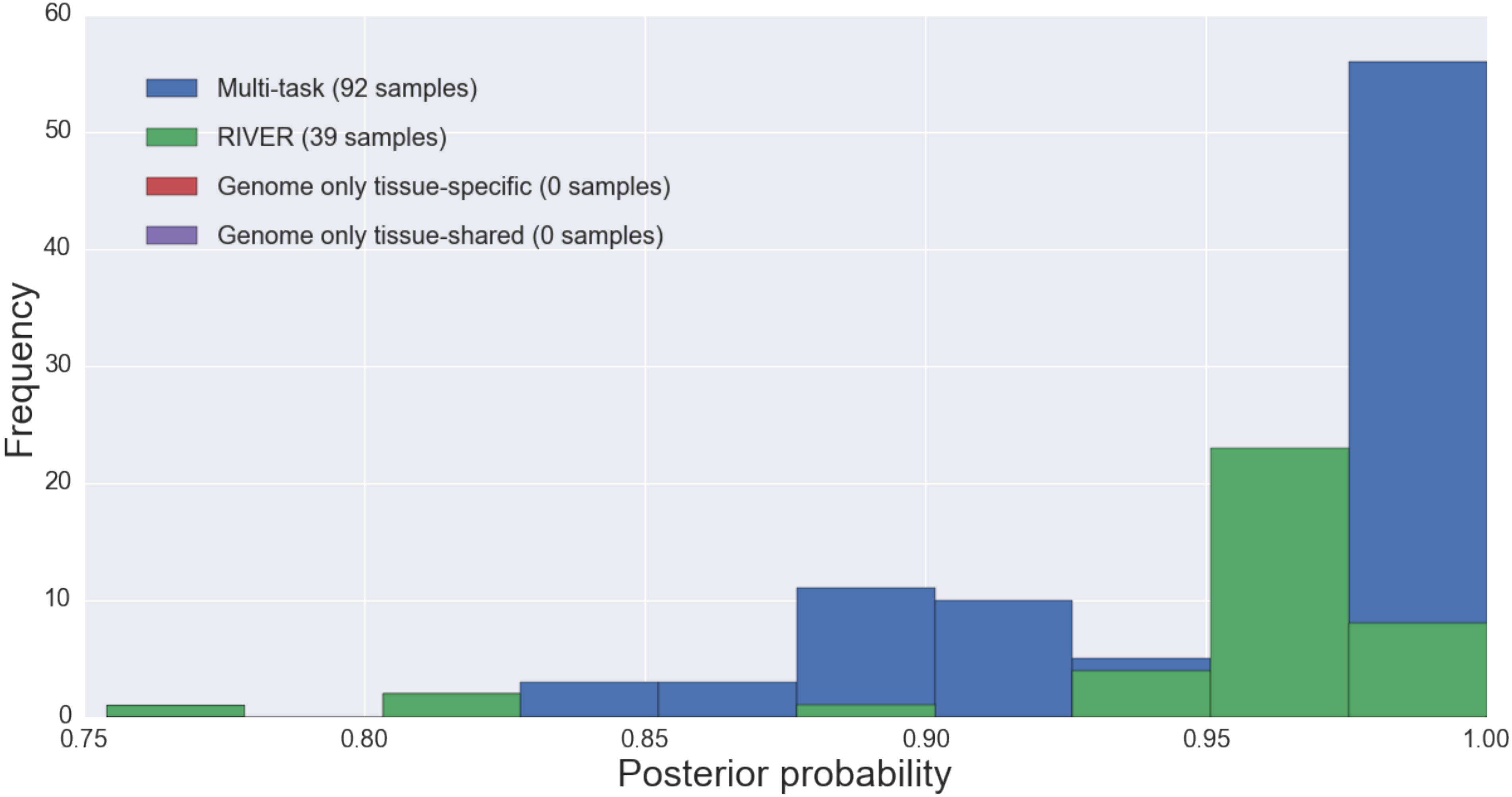
Presents an opportunity for researchers with WGS and (limited) RNA-seq to reliably identify functional rare variants

Thank you!

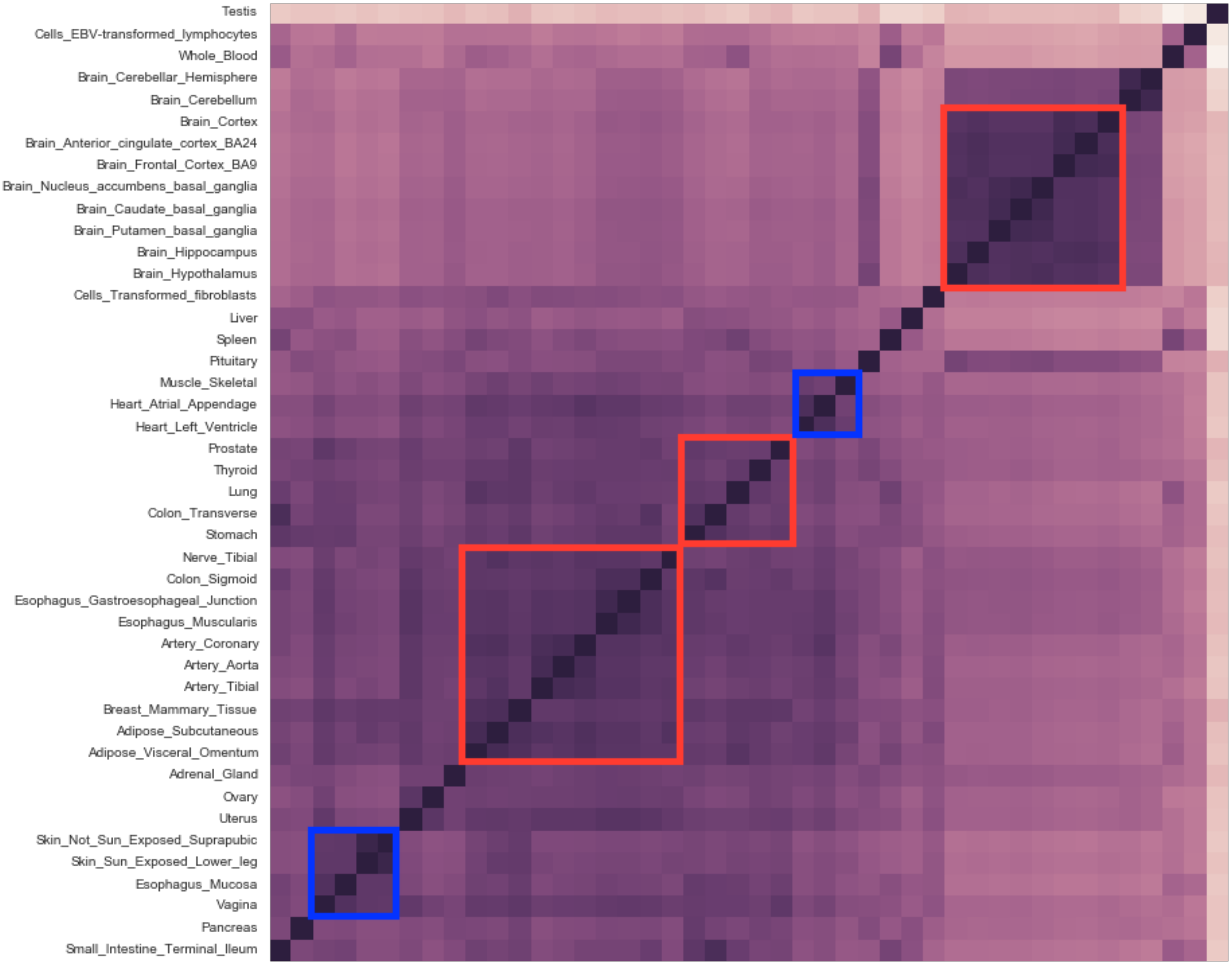
Battle Lab
Yungil Kim
Ben Strober
Alexis Battle

Montgomery Lab
Xin Li
Joe Davis
Emily Tsang
Zachary Zappala
Stephen Montgomery

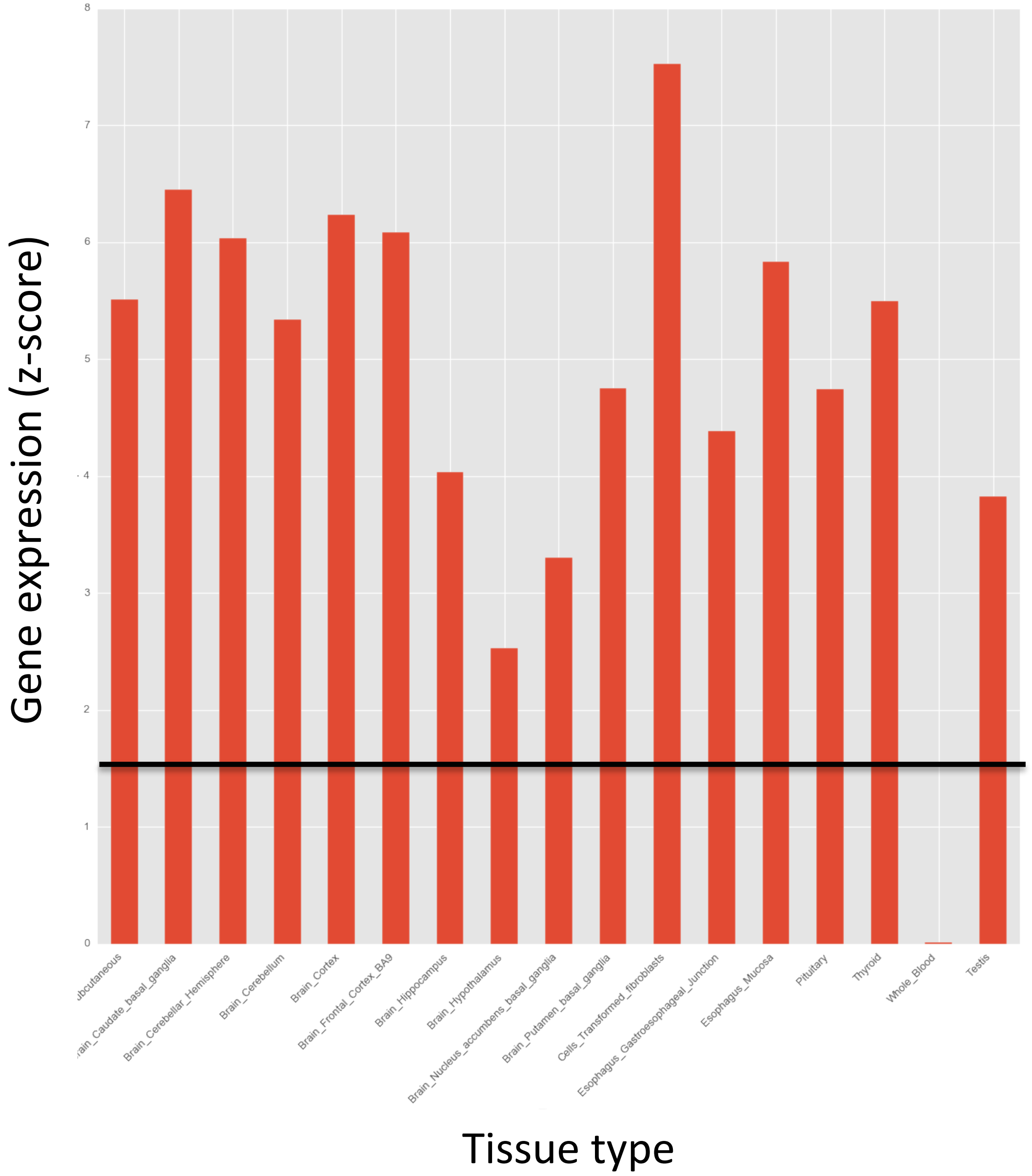
GTEx Consortium
Pistritto Fellowship
NIH
NIMH
Searle Scholar Program


Multi-task (92 samples)
RIVER (39 samples)
Genome only tissue-specific (0 samples)
Genome only tissue-shared (0 samples)
Frequency
0
10
20
30
40
50
60
0.75
0.80
0.85
0.90
0.95
1.00
Posterior probability

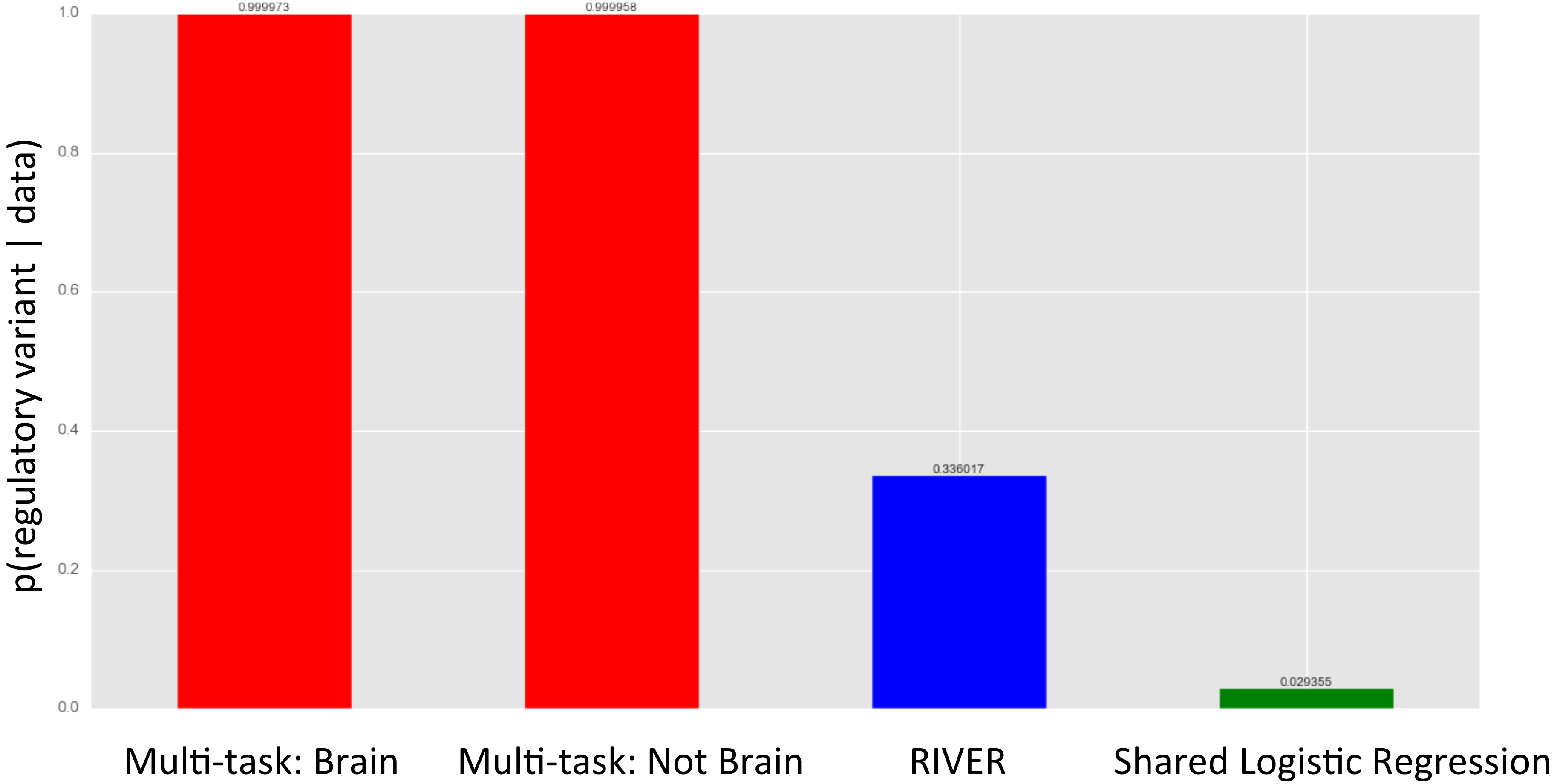
Tissue groups with similar behavior



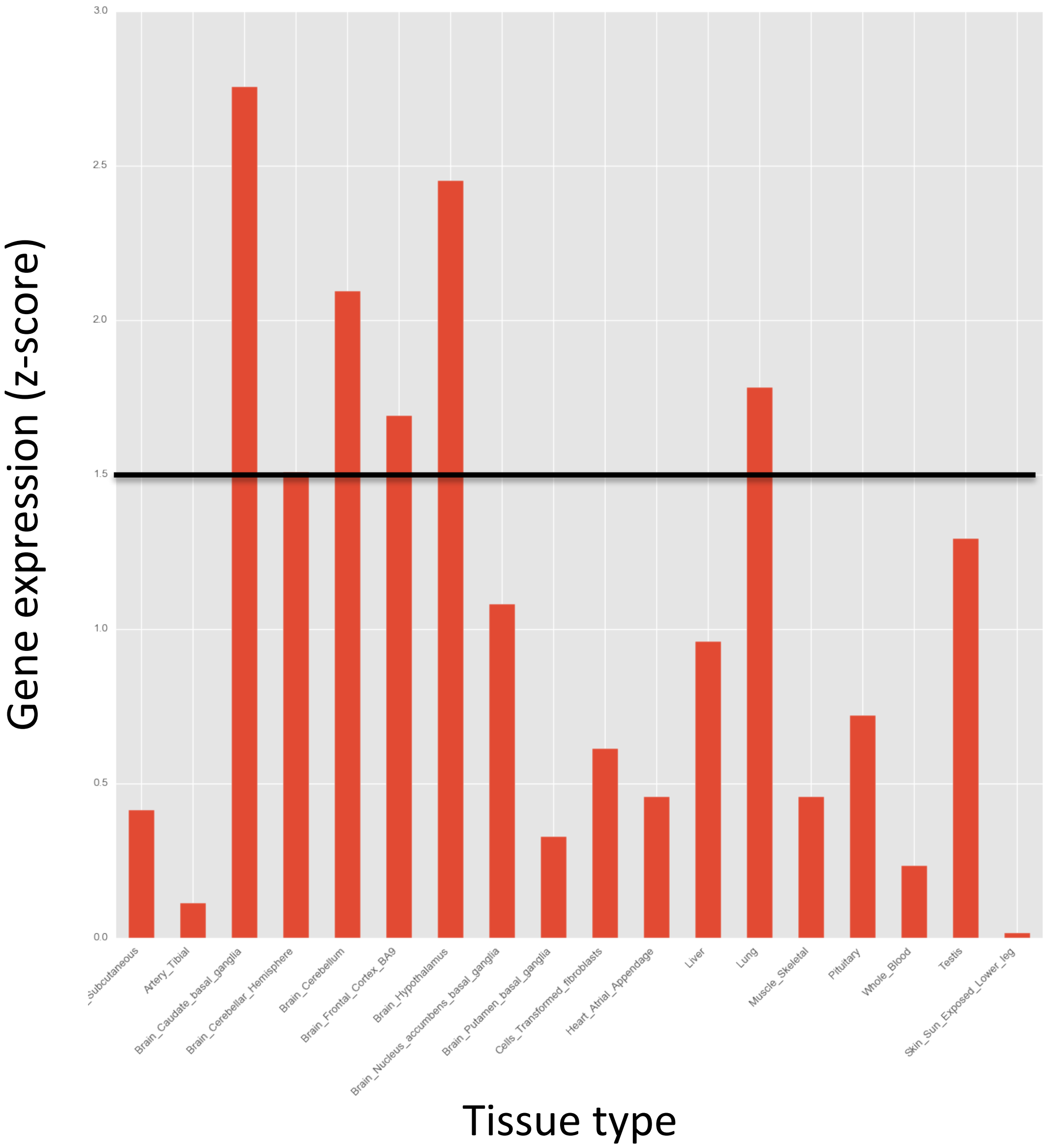
Case 1: Extreme expression across tissues



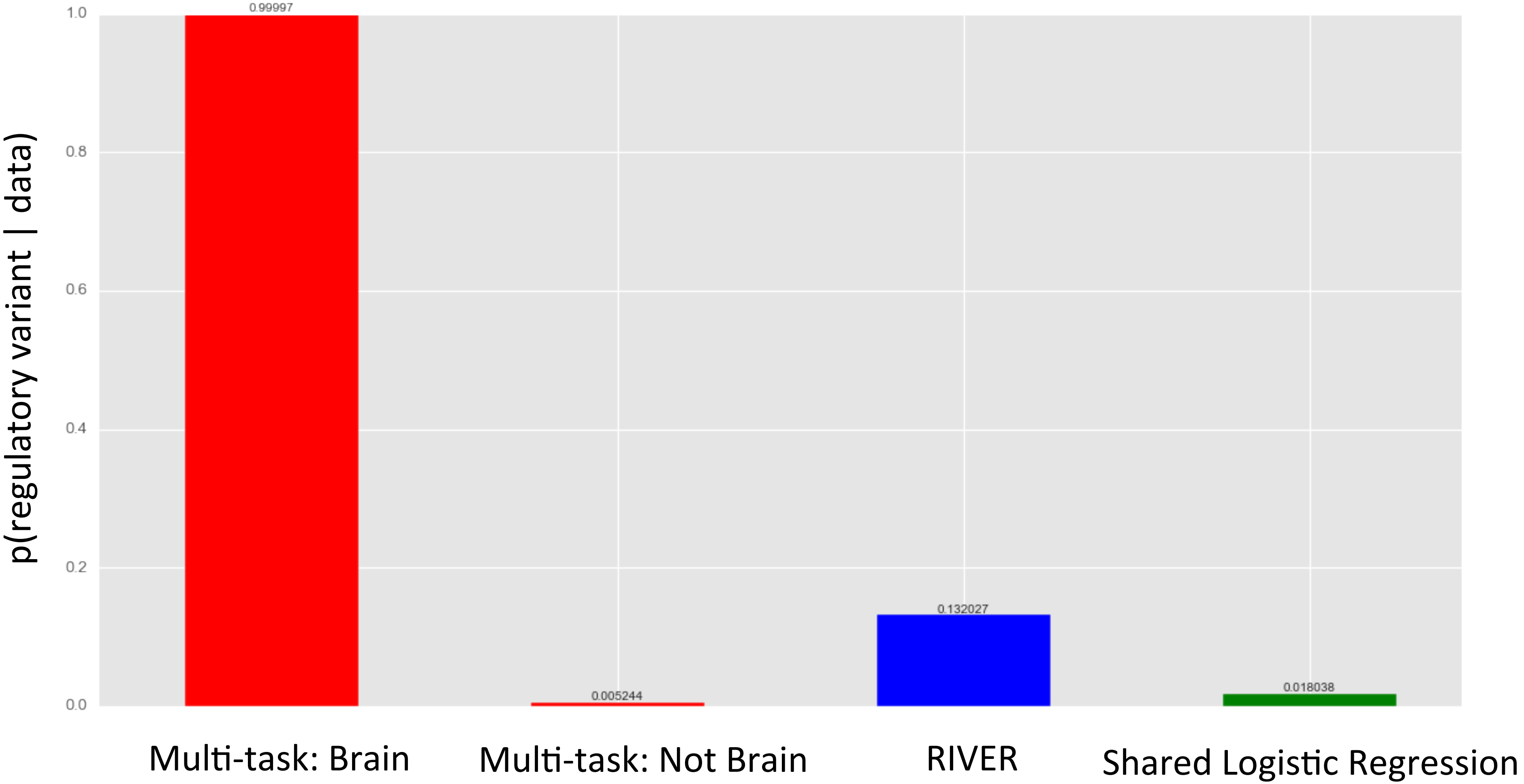
Model predictions



Case 2: Extreme expression in brain tissues



Model predictions


p(regulatory variant | data)
1.0
0.8
0.6
0.4
0.2
0.0
0.99997
0.005244
0.132027
0.018038
Multi-task: Brain
Multi-task: Not Brain
RIVER
Shared Logistic Regression